DOCUMENT RESUME

ED 320 938 TM 015 186

AUTHOR Badger, Elizabeth
TITLE Using Different Spectacles To Look at Student Achievement: Implications for Theory and Practice.
PUB DATE Apr 90
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Comparative Testing; Grade 12; High Schools; *High School Students; *Mathematics Tests; *Multiple Choice Tests; *Science Tests; Secondary School Mathematics; Secondary School Science; *Test Format; Theory Practice Relationship
IDENTIFIERS *Open Ended Questions

ABSTRACT

Data from science and mathematics tests administered to 12th graders in Massachusetts were analyzed to compare the use of multiple-choice items with that of open-ended questions. Specifically, the study was designed to determine the statistical equivalence of the two types of test items, the information yielded by the items, and the grounds upon which one should be preferred over the other. Mathematics tests included 380 multiple-choice items and 12 free response questions, while science tests included 361 multiple-choice items and 11 free response questions. Data were taken from test administrations to students attending 27 schools randomly selected from a total population of 297 schools participating in the study. The number of students in each school ranged from 190 to 651, with a median of 15 students completing each form. A total of 100 students completed each open-ended question. Results indicate that students who perform well when presented with multiple-choice questions perform at comparable levels when presented with more open-ended questions; however, they do not perform equally well. Data from use of the open-ended questions provide far more direct evidence about students' understanding than do data from use of the multiple-choice items. Finally, open-ended responses show potential as instructional devices. Six sample test items and seven tables are included. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

M. E. BADGER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED320938

Using different spectacles to look at student achievement: Implications for theory and practice.

Elizabeth Badger, Ph.D.
Massachusetts Department of Education

American Educational Research Association. Annual meeting.
April 1990
Boston, Massachusetts

BEST COPY AVAILABLE

TM015186

Using different spectacles to look at student achievement: Implications for theory and practice.

Elizabeth Badger
Massachusetts Department of Education

Background:

Evidence that U.S. student achievement in the areas of science and mathematics is lagging behind that of other nations, (La Pointe, et al. 1989; McKnight el al. 1987) has provoked significant responses from two sectors of the community. Policymakers at state and federal level make up one group; mathematics and science educators make up the other. Unfortunately, although there is general agreement on the extent of the problem, the solutions advocated by these two groups are not necessarily compatible.

Policymakers' attempts at reform have centered around issues of accountability. They have assumed that publicity engendered by test results will galvanize schools to reform their educational practices. In their turn, educators have turned to a close analysis of the content itself in their efforts at remediation. Publications such as the **Curriculum and Evaluation Standards** (National Council of Teachers of Mathematics, 1989) and **Science for All Americans** (American Association for the Advancement of Science, 1989) have chosen to ignore issues of accountability, stressing instead the instructional requirements demanded by the subject itself.

The incompatibility between the two approaches stems in large part from the different epistemological assumptions of each group, as well as their differing paradigms for measuring achievement. For policymakers, knowledge is defined as information. Arguing in terms of mathematical or scientific "literacy," public officials refer to the skills and factual knowledge that are needed in technological society. In contrast, when content experts speak of knowledge, they are more inclined to stress individual relevance and understanding, as well as the essentially tentative nature of theories and the social context of concepts.

These differing perspectives on what knowledge consists of have led to different paradigms for measuring success. If knowledge is conceived primarily as information, an individual data base that is available for use, then it can be quantified in terms of a single indicator. The prime

concern becomes, how can we develop a reliable and objective instrument for measuring what a person knows? In this case, it is logical for policymakers to turn to psychometricians for a model. If, on the other hand, knowledge is conceived as complex and personal, summary measures will not do. In order to validly measure the kinds of understanding that the subject demands, one must turn to the areas of developmental and cognitive psychology for the measurement paradigm. Essentially, this is what educators have done.

As a result of this basic conflict in purpose and vision, both sides are frustrated. The testing that is currently mandated by policymakers does not address itself to the kind of curriculum and instruction that is advocated by mathematics and science educators. Instead of assessment tasks that require an integration and application of mathematical and scientific concepts, in the majority of cases students are presented with a series of highly structured items that primarily require the retrieval of knowledge. On the other hand, curriculum experts have not offered alternative measures that satisfy the legitimate demand for public accountability. At the least, this misalignment between testing and curriculum casts doubt on the validity of current measures of mathematics and science achievement. More significantly, however, it makes any kind of curricular reform very difficult, if not impossible, for the teacher.

This study examines the characteristics of test items that exemplify the two approaches to measurement. It contrasts the results obtained on a set of multiple-choice items with those achieved on a set of open-ended questions that allowed for free response. The multiple-choice items were selected from a large battery of items designed to compare mathematics and science achievement among Massachusetts schools. The open-ended questions formed part of the same assessment, but were designed to give information on students' understanding an application of mathematical and scientific concepts. Scores on the two types of tests were compared in order to examine whether or not they measured the same construct. Individual items of each type were then examined for the quality of information provided.

The basic questions that the investigation attempts to answer are: Are these two types of test items statistically equivalent? Do they yield similar information? On what grounds should one be preferred over the other?

The Study:

In January 1990, an assessment was administered to all twelfth grade students in Massachusetts. In the area of mathematics, the test included 380 multiple choice items and

12 free response question. In science, the test included 361 multiple choice items and 11 free response questions. Matrix-sampling was used to distribute the multiple-choice items among 20 forms. Nineteen forms contained 20 science and 20 mathematics items. The twentieth form contained 10 mathematics items, 10 science items, as well as a section containing open-ended questions in either science or mathematics. The open-ended questions were also matrix sampled. Each of four forms contained 7 questions regarding either science or mathematics. (Social studies and reading were also included in the assessment and were paired with science and mathematics, respectively.)

The Sample. The design of assessment allowed for school comparisons across forms and for individual comparisons between the multiple-choice and open-ended responses in Form 20. Accordingly, 27 schools were randomly selected from a total population of 297 schools. The number of students in each of these schools ranged from 190 to 651, with a median of 15 students completing each form. Since the open-ended questions were spiralled in Form 20, a total of 100 students completed each open-ended question (from 2 to 10 in each school).

Question One: Is Form 20 a valid indicator of the math and science achievement as measured by the assessment as a whole?

The mathematics and science scores by which schools are rated are based on the aggregated scores of the 20 forms. However, only one form of the test, Form 20, contained both multiple-choice and open-ended items. Furthermore, the number of science and mathematics multiple-choice items in Form 20 was half the number contained in the other forms. Consequently, in order to generalize from any comparisons of students' performance on the two types of items, it was necessary to establish the validity of Form 20 as an indicator of achievement.

School results for Form 20 were correlated with results for the other 19 forms of the assessment and the total math/science score across all forms. Despite its relative brevity, schools' science score on Form 20 was moderately to highly associated with the scores derived from the other forms. The correlations, ranging from 0.47 to 0.84, were comparable in size to the inter-correlations among scores on the other forms and significant at the .01 level or greater. Furthermore, the correlation between the science score on Form 20 and the total science score was 0.77, compared to a median correlation of 0.79 between the other forms and total test score.

In mathematics the correlations between Form 20 and the other forms ranged from 0.62 to 0.87. The correlation between the mathematics score on Form 20 and the total mathematics score was 0.87, compared to a median correlation of 0.82 between the other forms and the total test score.

Conclusion. The high correlations between Form 20 and a more comprehensive measure consisting of hundreds of items indicates that, despite its relative brevity, Form 20 can be considered a statistically valid indicator of the math and science test scores achieved by the schools.

Question Two: Can responses to the open-ended items predict performance on the multiple-choice items, and vice versa?

In order to examine this question, the multiple-choice and the open-ended responses given by same students were compared. Although all students who took Form 20 of the assessment completed the same multiple-choice items (10 in each subject area), the open-ended questions were spiraled among four booklets. The number of students completing each booklet is as follows:

booklet	# Items	# Students
Form 20A Mathematics	7	92
Form 20B Mathematics	8	98
Form 20A Science	7	73
Form 20B Science	4	91

Scores of students who took each form of the open-ended questions were correlated to their multiple-choice results in the same subject area with the following results.

	Mathematics Open-Ended	
	Form A	Form B
Multiple-Choice	0.42	0.60

	Science Open-Ended	
	Form A	Form B
Multiple-Choice	0.49	0.44

The correlation between the two types of formats are moderately high. The percentage of the variance in one variable that was shared by the other ranged from 17 to 36

measures based on single items or sets of items with small numbers of items in them. (See discussion in footnote 1.) This should be kept in mind when interpreting the results.

The percentage of the variance in one variable that was shared by the other ranged from 17 to 36 percent. Furthermore, when the relationship of individual items to total score was examined, it was apparent that the correlations of open-ended items to the total multiple-choice score was generally higher than the inter-correlation of the multiple-choice items themselves. (See Appendix.) For example, among the multiple-choice mathematics items, inter-correlations ranged from -0.02 to 0.36 with a median of 0.19, while the median correlations between the open-ended responses and the total multiple-choice score was 0.43 and 0.38 for forms A and B, respectively. Inter-correlations among the multiple-choice science items ranged from -0.05 to 0.22 with a median of 0.09, while the correlations between the open-ended science responses and the total multiple-choice score were 0.34 and 0.36 for forms A and B, respectively.

Conclusion: Since correlation indicates the constancy of relative position among sets of scores, it may be concluded that, whether students are tested via multiple-choice items or open-ended ones, their relative performance does not differ significantly. If we are interested only in rank scores for students or schools, the former can serve as a sufficiently accurate and efficient indicator of relative achievement.

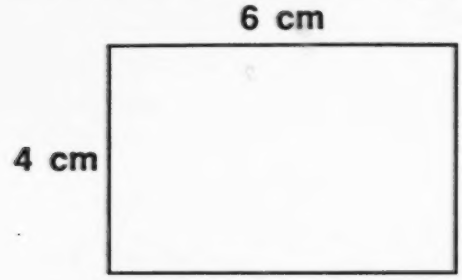
Question Three: Over the test as a whole, how do items that ostensibly measure the same topic differ in terms of the information provided?

Because there were too few multiple-choice items in Form 20 to cover a topic in any depth, items for this analysis were selected from all forms of the assessment. The assumption was made that the random distribution of the forms, as well as their comparable difficulty, allowed for the generalization of results.

MATHEMATICS

The topic of **area** was chosen as the focus for comparisons in mathematics. There were 12 items concerning area in multiple-choice format and one open-ended question. The four multiple-choice items discussed represent the typical range of difficulty and conceptual challenge.

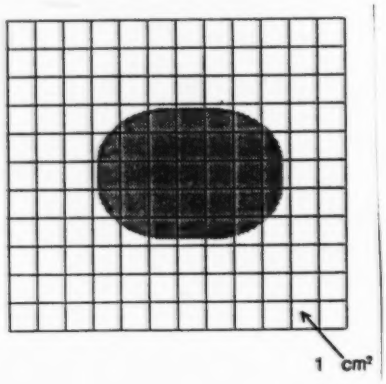
A. What is the area of the figure below?



p = .76

This item represents the simplest type of question regarding area. Students are required merely to compute the figures given for length and width. The incorrect options presented, which include the correct figure for the perimeter, signal whether the students can associate the correct operation with the word <u>area.</u>

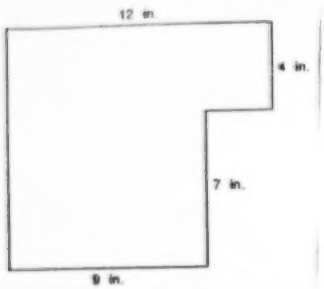
B. About how many small square cm in the shaded region?



p = .62

This item measures the ability of students to approximate area, given a grid. Incorrect options include the measurement of all units with any shaded area and units with only completed shaded areas. For students who regard area only in terms of the measurement of rectangles, this may be a novel problem. Although an understanding of approximation of area is important to the process of integration, students' success may depend completely on what units they considered as shaded.

C. What is the area of this figure?



p = .46

This item calls upon students' ability to visualize a completed figure, as well as their facility at computation. Before beginning to calculate the actual area of the figure, they must be able to deduce the measurements of the missing two sides. Incorrect options include the total measurement of all sides (perimeter), the total measurement of the given sides (perimeter), and the product of the various combinations provided in the diagram. Although a student's selection of an <u>incorrect</u> option may give information on the extent to which they confuse the terms **area** and **perimeter**, the selection of a <u>correct</u> option does not tell us about students' ability to partition or to visualize a completed figure.

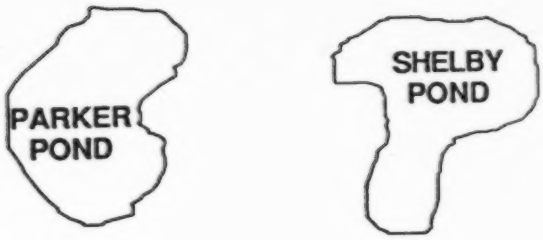
D. One pound of grass seed will cover an area of 400 square feet. How many pounds of grass seed are needed to cover a rectangular yard that is 120 feet long and 90 feet wide?

p = .59

This problem calls for two operations. First, students must recognize and calculate the area; then they must recognize and calculate the proportional relationship. Despite the fact that the calculation of area was but the first step in solution, almost 60 percent of the students chose the correct option. Incorrect options included calculation errors, the sum of the two dimensions given (210), and that sum divided by 40. The difference in success rate in this problem and the one above (Problem C) suggests that students may perform better when problems are set in a a practical context.

In addition to the above items, which appeared on the multiple-choice sections of the assessment, the following question was included in the open-ended response section of Form 20.

E.



Theo wants to find out which pond covers the larger area, Parker Pond or Shelby Pond. He does not need to know the two areas---just which is larger. Theo claims that all he has to do is measure the distance around each pond to find out what he wants. Will his method work? Use the space below to make a convincing argument for your answer.

Results:

3 points: Discusses difference between perimeter and area and supplies counter-example. 9%

2 points: Gives counter-example and shows a recognition of difference between area and perimeter, but not as complete as above. 7%

1 point: Answer is inarticulate but implies correct explanation 6%

0 points: Incorrect answer or blank 76%

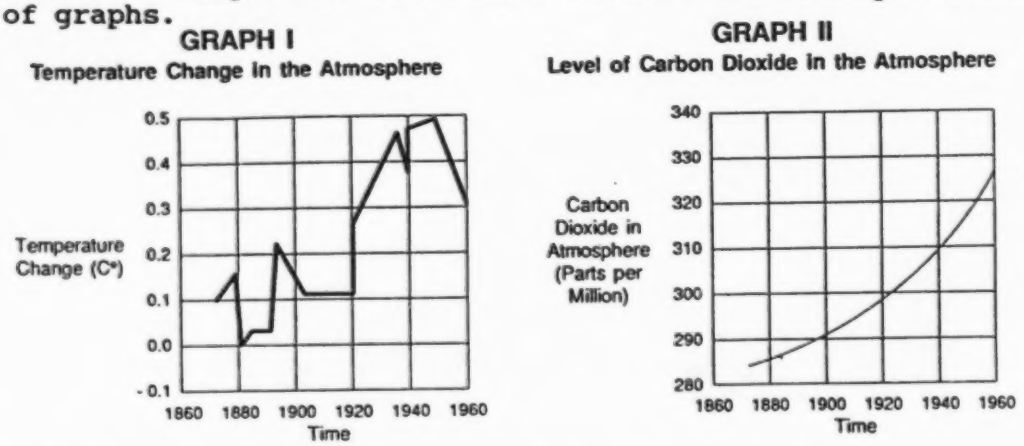
It is not surprising that this, more challenging question, received far fewer correct responses than did the multiple-choice items discussed above. Approximately 35 percent of the students answered yes, Theo could find the area by measuring the distance around the pond; while 8 percent answered that he would have to knows the depth of the pond or its circumference; and 5 percent implied that it was impossible to ascertain the area of figures other than rectangles.

The question taps students' understanding of what <u>area</u> means. It does not ask for formula, nor does it include the ubiquitous rectangle. Rather it requires students to examine the relationship between linear and surface measurement. In addition to the conceptual demands of the question, it also taps students' ability to "talk mathematics". In this case, the demands are informal and do not depend upon algebraic notation. However, the level of students' dialogue is also measured and points are awarded for an appropriate counter-

example. (See Appendix for scoring guide and categorization of responses.) Furthermore, there is a clear tie-in with instruction. This type of task suggests a number of instructional activities, as well as discussion topics, through which students can explore the meaning of area (Stimpson, 1989).

SCIENCE

A similar comparison in science involves the interpretation of graphs.



Using the above graphs, the students are asked to identify the **best** interpretation that could be made about the changes in temperature and carbon dioxide levels between 1870 and 1955. The options are:

a. The amount of carbon dioxide has increased steadily and the temperature shows an overall increase.

b. The amount of carbon dioxide and the temperature have increased steadily.

c. The amount of carbon dioxide has decreased steadily and the temperature shows an overall decrease.

d. The amount of carbon dioxide has increased steadily causing a varying change in temperature.

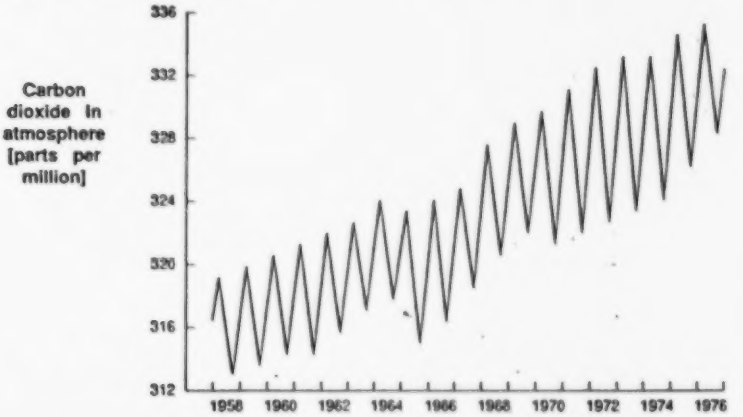
This question is designed to measure students' ability to integrate information contained in graphical material in order to make reasonable inferences. It is assumed that the choice of options b or c would indicate a misreading of the graphs. The selection of option d would indication that students inferred a causal relationship that is not

warranted by the data presented. This last option appeared convincing enough to attract almost half of the students. The results were as follows:

	response frequency
option a	37
option b	8
option c	6
option d	49

In a similar open-ended question students were presented with the following graph and were asked 1) to give reasons for the general increase in carbon dioxide over the time periods portrayed; and, 2) to account for the fluctuations.

The graph below shows the concentration of carbon dioxide in the atmosphere over a period of almost two decades.



Students performed well in response to the first question, "What reasons can you give for the general increase?". Thirty-four percent gave a clear account of possible reasons for the increase, describing the relationship of carbon dioxide production to such events as the burning fossil fuels, increased industrialization, and the loss of natural vegetation. Another 35 percent gave a less detailed but adequate explanation. These results are consistent with the results achieved on multiple-choice question that measured students' knowledge about the production of carbon dioxide. For example, when asked "Where does carbon dioxide come from?" with reference to humans, 54 percent chose the option, "It is a waste product from breaking down food for energy."

Students performed poorly, however, in response to the second question. In this case, 75% of the students were unable to account for the fluctuations. Few appeared to notice that there were fluctuation, and 35 percent of the

students ignored all but the general trend showing that carbon dioxide had risen over the period.

However, the distinction between these two types of items lies less in the success rate of the students than it does in the quality of the information gained. In the multiple-choice item, inferences must be drawn about students' thinking. The test developer assumes that the behavior which the options are designed to measure is the behavior elicited from the student. In other words, when the student choses option d, we infer that s/he misunderstands the notion of causality, but there is no direct evidence that this is the case. Indeed, an equally plausible assumption may be that the student is not even examining the graph but, instead, responding only to the accuracy of the statement. In contrast, the open-ended question gives direct evidence about how the student has read the graph. When students' answers focused only on the general trend of the data (which confirmed their prior knowledge), we know that they have ignored irregularities that appear to be disconfirming or irrelevant.

Conclusion. It appears that the two types of questions in mathematics and science differ in both difficulty and in the kinds of information that can be obtained. Students found open-ended questions more difficult to answer. However, the information obtained from the open-ended gave clearer evidence of their understanding of concepts and well as their intellectual processes.

Summary

When relative performance among individual students is compared, the results appear to be similar. The correlations between the summary scores indicate that students who perform well when presented with questions in multiple-choice format perform at comparable levels when presented with more open-ended questions. They do not, however, perform equally well. When asked to justify, explain, extrapolate or infer, fewer are able to succeed without the benefit of some probable answers to select from.

Secondly, the examination of individual items indicated that the open-ended questions gave far more direct evidence about students' understanding than did the multiple-choice. No matter how carefully the options were chosen, the type of misunderstanding or lack of skill on the part of students could only be inferred from the incorrect options chosen. Single multiple-choice items were incapable of yielding significant information about students' understanding of the topic. Inferences about students' understanding could be based only on information gathered from a collection of

items regarding the same topic. In contrast, when students were presented with an open-ended question that required conceptual understanding, as in the case of area, their misconceptions were immediately evident. Furthermore, when two analogous science items were contrasted, the influence of extraneous ideas (such as the link between carbon dioxide and heat) was apparent in the open-ended responses. In the case of the multiple-choice item, inferences about the sources of error had to be based on the selection of options and were tenuous at best.

Thirdly, analysis of the open-ended responses indicates their potential as instructional devices. Because multiple-choice items are designed to produce a single correct response, they are not conducive to the investigations or discussions that promote learning. In contrast, the generative response that is required by open-ended questions leads to a greater range of acceptable answers and strategies, as well as providing useful evidence of misconceptions and faulty reasoning.

Implications for theory and practice.

The evidence suggests that, when the purpose of testing is limited to producing an indicator of general mathematics or science achievement, a battery of multiple-choice items is an appropriate and sufficient measure. It is an efficient form of gathering a wide range of data about school performance, particularly when it is administered in matrix form. It acts as a useful discriminator of student achievement within a well-defined domain. Furthermore, scores from more time-consuming and expensive forms of assessment, such as open-ended tasks, do not yield significantly different results regarding the relative achievement of students. Students who perform well on a multiple-choice are likely to perform well on open-ended questions, and vice versa.

These results should not be considered surprising, given the large degree to which content knowledge has been found to contribute to the ability to think clearly and inventively in a subject area. Although multiple-choice items may be primarily limited to the recall of factual knowledge or the application of skills, there is evidence that such knowledge and skills are essential components of more skilled and thoughtful performance. That the multiple-choice scores correlate well with the more demanding tasks that were presented in the open-ended form of the test is evidence of their test validity.

On the other hand, whatever their ostensible intent, tests are regarded as more than indicators of general achievement.

They do not exist in a vacuum, but exert powerful influences on what is taught and what is learned. For this reason, there are other aspects of test validity that must be considered. Variously termed instructional validity (NCTM, 1989), ecological validity (Murnane) and systemic validity (Frederickson and Collins, 1989), these aspects of validity address the extent to which tests reflect what is or should be taking place in the classroom. They do not focus on the statistical relationship of scores, but on the content of the items themselves. When examined in these terms, the open-ended tasks are more reflective of the types of behavior that is the object of education, are more conducive to the kinds of teaching that educators advocate, and provide more useful information to both students and their teachers.

Multiple-choice items may be effective discriminators among strong and weak students; however, when the mandate extends to what students can do, how they understand, and how they can apply their knowledge, more ecologically valid testing is called for.

---

footnote 1, page 5: The magnitude of these correlations should be judged with reference to the units involved. When results on a single test item constitute one or both of the measures being correlated, large values of correlation coeeficients are not expected. Item-to-item correlations would be quite low: item-to-scale correlations somewhat higher. When scores on sets of items are involved, correlation should be higher--how high depends largely on the number of items in a set. This is consistent with what we know about the relationship between test length (number of items) and test reliability.

In selecting items for inclusion in a test, point-biserial correlations are often used as one of several indicators of item quality. A point-biserial correlation is simply the correlation between an item (scored correct/incorrect or 1/0) and the total score on a set of items (i.e., item - to-scale correlations). For tests consisting of a set of fairly hetrogeneous items, point biserial correlations as low as 0.2 could be quite reasonable.

I wish to acknowledge the assistance of Beverly Miyares of the Massachusetts Department of Education and Charles DePascale of Advanced Systems for their assistance in the computer analysis of the data.

References.

American Association for the Advancement of Science. 1989. Science for All Americans. A Project 2061 Report. Washington, D.C.: AAAS.

Frederiksen, J.R. and A. Collins. 1989. "A systems approach to educational testing." Educational Researcher, vol. 18, 9, 27-32.

LaPointe, et al. 1989. A World of Differences. Princeton, N.J.: Educational Testing Service.

McKnight, C.C. et al. 1987. The Underachieving Curriculum: Assessing U.S. School Mathematics from an International Perspective. Champaign, Ill.: Stipes Publishing Co.

National Council of Teachers of Mathematics. 1989. Curriculum and Evaluation Standards for School Mathematics. Reston, VA: NCTM.

Perkins, D.N. and G. Salomon. 1989. "Are cognitive skills context-bound?" Educational Researcher, 18, 1, 16-25.

Schoenfeld, A.H. 1985. Mathematical Problem Solving. New York: Academic Press.

Stimpson, V.C. "What do we mean by area and perimeter?", Mathematics Teacher, vol. 82, 5,1989, 342-344.

APPENDIX

MATHEMATICS: CORRELATIONS BETWEEN MULTIPLE-CHOICE ITEMS AND SCORES ON OPEN-ENDED FORMS OF THE ASSESSMENT

	Open-ended Scores	
m-c items	Form A (n=92)	Form B (n=98)
1	0.19	0.35
2	0.35	0.48
3	0.22	0.40
4	0.11	0.31
5	0.26	0.37
6	0.28	0.22
7	0.03	- 0.14
8	0.28	0.35
9	0.07	0.20
10	0.06	0.41

MATHEMATICS: CORRELATIONS BETWEEN OPEN-ENDED ITEMS AND SCORE ON THE MULTIPLE-CHOICE SECTION OF THE ASSESSMENT

Form A Items	Multiple-choice scores
1	0.43
2	0.44
3	0.44
4	0.38
5	0.43
6	0.26
7	0.05

Form B Items	
1	0.37
2	0.50
3	0.38
4	0.35
5	0.38
6	0.39
7	0.36
8	0.44

SCIENCE: CORRELATIONS BETWEEN MULTIPLE-CHOICE ITEMS AND SCORES ON OPEN-ENDED FORMS OF THE ASSESSMENT

	Open-ended Scores	
m-c items	Form A (N=73)	Form (N=91)
1	0.04	0.24
2	0.32	0.26
3	0.30	0.12
4	0.19	0.17
5	0.15	0.28
6	0.03	0.01
7	0.35	0.12
8	0.14	0.17
9	0.37	0.17
10	0.19	0.20

Open-Ended Items	Multiple-Choice Score
Form A	
1	0.36
2	0.14
3	0.31
4	0.26
5	0.34
6	0.35
7	0.39
Form B	
1	0.26
2	0.36
3	0.35
4	0.42